Master's Thesis

Zero-Shot Semantic Segmentation via Spatial and Multi-Scale Aware Visual Class Embedding

Sungguk Cha

Department of Computer Science and Engineering

Graduate School of UNIST

2021

Zero-Shot Semantic Segmentation via Spatial and Multi-Scale Aware Visual Class Embedding

Sungguk Cha

Department of Computer Science and Engineering

Graduate School of UNIST

Zero-Shot Semantic Segmentation via Spatial and Multi-Scale Aware Visual Class Embedding

A thesis submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Sungguk Cha

11. 27. 2020

Approved by

Adviser
Kwang In Kim

Zero-Shot Semantic Segmentation via Spatial and Multi-Scale Aware Visual Class Embedding

Sungguk Cha

This certifies that the thesis of Sungguk Cha is approved.

11. 27. 2020

signature

Adviser: Kwang In Kim

signature

Seungryul Baek: Thesis Committee Member #1

signature

Sung Whan Yoon: Thesis Committee Member #2

Abstract

As a cost effective learning, word based zero-shot semantic segmentation (w-ZSSS) approaches are proposed, which recognizes an unseen target class only with a word vector and without a supporting image. The expressiveness of w-ZSSS is limited because their class representation of a novel class is constant. Tackling w-ZSSS, we propose a Spatial and Multi-scale aware Visual Class Embedding Network (SM-VCENet) for zero-shot semantic segmentation. SM-VCENet generates visual class embedding of an unseen class by transferring visual context knowledge on the query image, resulting domain-aware class representation. SM-VCENet enriches visual information of visual class embedding by incorporating multi-scale attention and spatial attention. Our SM-VCENet outperforms the state-of-the-art with a noticeable margin on the PASCAL and COCO test sets. We also propose a novel benchmark (PASCAL2COCO) for zero-shot semantic segmentation, which includes domain adaptation and more challenging samples.

Contents

List of Figures

List of Tables

CHAPTER I

Introduction

By the advent of convolutional neural networks, previous methodologies for semantic segmentation have achieved super-human accuracy [1–5] and speed [6–10] on the benchmarks such as MS COCO [11] and PASCAL VOC [12]. However, aforementioned methods are impractical in the real-world because they require expensive costs for annotation of large-scale dataset and fail to predict a novel class that is unseen during the training phase. Overcoming them, the proposed **Z**ero-**S**hot and **F**ew-**S**hot **S**emantic **S**egmentation (**ZSSS** and **FSSS** respectively) approaches materialize class representative embedding of an unseen target class (hereinafter, class embedding) with zero or few number of target class image(s) and recognize the novel class by comparing the class embedding and query image feature (shown in Figure 1.1 (c) and (d)). In detail, the recently proposed word vector based ZSSS (w-ZSSS) approaches [13–15] materialize the class embedding of a novel class with word vector, such as GloVe [16] that covers any class name and contains semantic information in language domain, while FSSS approaches [17–22] generate the class embedding from a few number of target class images.

Zero-shot learning approaches proposed to use a word vector for the class embedding, but no one showed *why we can use a word vector for an image recognition task.* The recently proposed w-ZSSS approaches (Kato *et al.* [14] and Yongqin and Subhabrata *et al.* [15]) even assumed the knowledge distributions of language and vision are the same, directly using the word vector as the class embedding. In contrast, we hypothesized *the two distributions are different, so we can solve the recognition problem better with only visual domain knowledge and without language domain knowledge.*

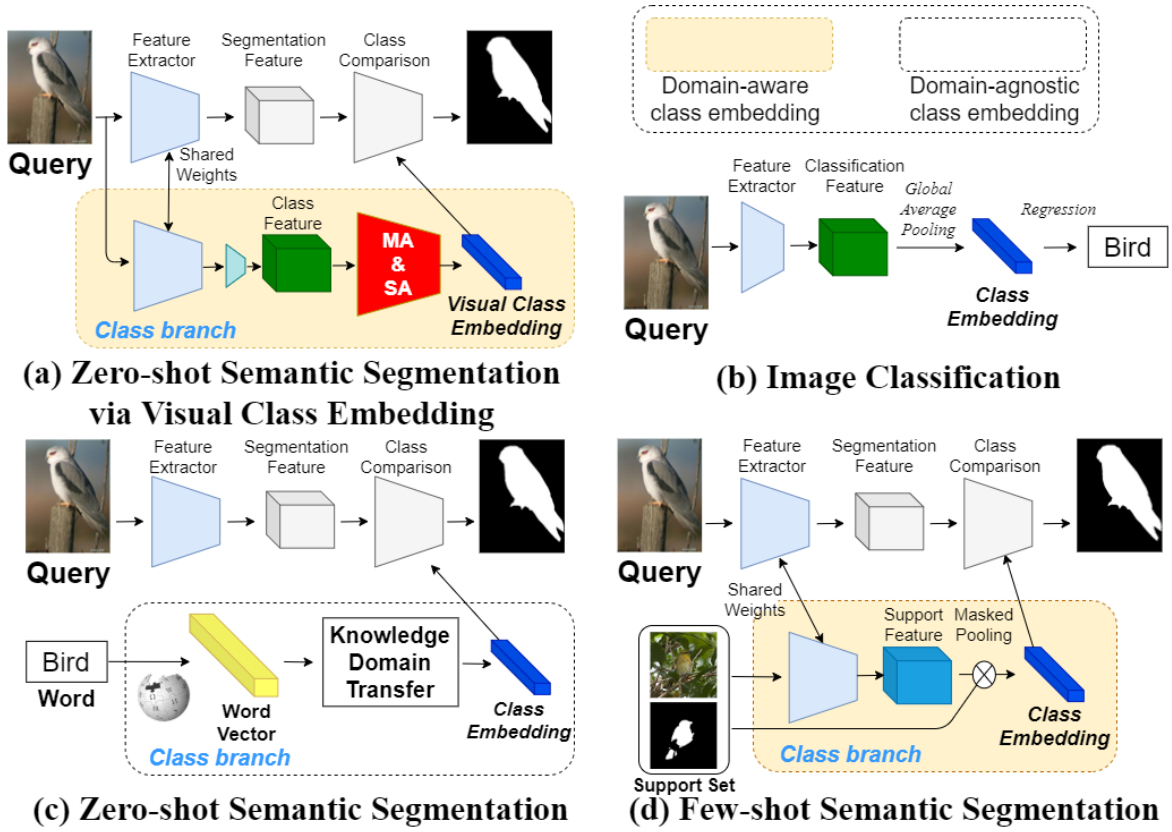

Figure 1.1: Zero-shot(few-shot) semantic segmentation approaches abstractions. Domain-aware means that the class embedding implies the query image distribution. MA & SA in (a) refers multi-scale attention and spatial attention. (a) Our proposed Visual Class Embedding is originated from the query image, so domain-aware. (d) Assuming the supporting images in the few-shot semantic segmentation follow the same distribution of the query image, the class embedding is domain-aware. (c) Word vector oriented class embedding is constant once trained and not flexible with respect to the domain distribution (domain agnostic).

In this paper, first, we present **S**patial & **M**ulti-scale aware **V**isual **C**lass **E**mbedding **Net**work (**SM-VCENet**). Our SM-VCENet generates domain-aware class embedding by transferring ImageNet [23] pretrained knowledge on the query image without any side information such as the word vector in w-ZSSS. SM-VCENet also incorporate spatial attention and multi-scale attention to classify pixels of multi-scaled and complex-structured objects. On the other hand, the class embedding in w-ZSSS approaches have limitations in the domain-awareness [24] and the representation of visual information. Compared to FSSS which generates the class embedding with supporting images that share the same distribution of the query image, w-ZSSS approaches have presented constant, domain-agnostic class embedding, depending only on the fixed word vector. Word vector originated class embedding cannot include any visual information such as multi-scale and spatial information. In contrast, our SM-VCENet preserves the query image domain information and enriches visual (spatial and multi-scale) information.

Moreover, we present a novel challenging benchmark (**PASCAL2COCO**) for ZSSS, containing domain adaptation problem and including multiple and complex objects in a noisy background. Previous ZSSS work [14] evaluated its performance on PASCAL-5^i [25] FSSS benchmark, where the unseen classes are in the same domain and most images contain a large single object. Our PASCAL2COCO benchmark can evaluate *generalization ability* by conducting domain adaptation in the different domain and contains more challenging samples, including *multi-scaled multiple objects* and *complex objects in noisy backgrounds*.

Our contributions.

1. We propose *domain-aware* SM-VCENet that achieves remarkable robustness in generalization for ZSSS, tackling domain-agnostic class embedding in w-ZSSS.

2. We show that our SM-VCENet generates class embedding capturing *multi-scale* and *spatial* information to recognize multi-scaled or complex objects in noisy backgrounds.

3. We first present ZSSS domain adaptation benchmark, in which a model is trained on PASCAL-5^i and evaluated on both PASCAL-5^i and COCO-20^i test set. In a sense that generalization is the most important idea of zero-shot learning, our proposed benchmark provides meaningful generalization measurement.

CHAPTER II

Related Works

2.1 Semantic Segmentation

Semantic segmentation requires multi-scale and spatial information understanding. In order to grasp multi-scale information, multi-scaled feature extraction and aggregation approaches [1, 26, 27] have been researched. Lowe [26] proposed scale-invariant feature transform (SIFT) that extracts a feature from multi-scaled images via Difference of Gaussian. Motivated, He *et al.* [27] and Zhao *et al.* [1] applied multi-scale feature extraction for modern deep convolutional neural networks(DCNNs). Spatial information understanding, including localization [28, 29] and global context information [30, 31], for DCNNs has been proposed. Long and Shelhamer *et al.*(FCN) [28] secured localization features by reducing resolution deductions in DCNNs. Further, Wang *et al.*(HRNet) [29] dramatically reduced the deduction by high-resolution convolutions, tackling the neighborhood constrained characteristic of convolution operation. Zhao and Zhang *et al.*(PSANet) [30] proposed point-wise spatial attention to relax the local neighborhood constraint. Recently, the nonlocal operation based networks [31–33] have been proposed to grap global context of features by enlarging the receptive field. Wang *et al.*(NNN) [31] first proposed the non-local operation that captures long-range dependencies by computing a output pixel value from weighted summation of all input feature values. Moreover, asymmetric [33] and compact-generalized [32] non-local network decreased complexity computation for the matrix multiplication by proposing asymmetric matrix dimension and compact representation for multiple kernels. In this work,

we adopted multi-scale feature extraction and non-local block [31] for *multi-scale* and *spatial* information aware class embedding.

2.2 Zero-(Few-)Shot Semantic Segmentation

Shown in the Figure 1.1, ZSSS and FSSS approaches share the framework, consisting of two major components: (1) class embedding generation (the class branch in Figure 1.1), and (2) class comparison between the segmentation feature and the class embedding.

First, FSSS [17, 18] and ZSSS [13, 14] approaches worked on generating expressive class embedding. Making class embedding to contain meaningful visual information in FSSS, Liu and Zhang *et al.*(PPNet) [17] proposed part-aware class embedding grouped by super pixel by SLIC [34]. In addition, Zhang *et al.*(PGNet) [18] proposed attention masked pooling. On the other hand, ZSSS that has no supporting visual source generates class embedding with language context knowledge, word vector. Bucher *et al.* [13] proposed ZS3Net that learns to generate synthetic features with word2vec and visual feature. Kato *et al.* [14] proposed ZSVM that maps the word vector into visual semantic space by variational sampling.

Second, comparing the class embedding and the segmentation feature is actively studied in FSSS [19–22] and ZSSS [15]. In FSSS, Siam *et al.*(AMP) [21] showed adaptively weighted classifier with the label masked pooling on support images. Liu *et al.*(CRNet) [19] and Wang *et al.*(PANet) [35] proposed query-support symmetric branch methods that mutually learn from both query image and support image. Zhang *et al.*(CANet) [20] presented the pixel-wise comparison framework with iterative optimization. Nguyen and Todorovic [22] proposed guided ensemble inference in the multi-shot setting. In the ZSSS setting, class comparison can be conducted by the same way as FSSS [14] or Xian and Choudhury *et al.*(SPNet) [15] proposed semantic projection, using the word vector as classifier weight directly.

In this work, we introduce a novel approach to generate class embedding even without side information such as the word vector in the ZSSS setting.

In addition, there is a meta-learning zero-shot domain adaptation work [36], requiring source domain and target domain samples. In contrast, our domain adaptation task does not require a target domain sample, aiming to *generalize* better.

CHAPTER III

Task Description

The goal of zero-shot semantic segmentation is to perform segmentation on a novel class that is unseen during the training phase. Here, we define the set of classes for training C_{train} and test C_{test}. From the global classes $C_{global} = \{c_1, c_2, ..., c_\infty\}$ that includes every class in the real-world, there are n training classes $C_{train} = \{c_1, c_2, ..., c_n\}$ and m target classes for test $C_{test} = \{c_{n+1}, c_{n+2}, ..., c_{n+m}\}$. Note that there is no intersecting class between the training classes and the test classes $C_{test} \cap C_{train} = \emptyset$.

3.1 ZSSS using word vector.

Recent zero-shot semantic segmentation methods [14, 15] utilizes word vectors for side information to generate class embedding. In the ZSSS setting using the word vector, the test set S^{test}_{word} contains tuples of query image and the corresponding word vector. The training set S^{train}_{word} consists of query image, label and corresponding word vector. Thus, the two sets S^{train}_{word}, S^{test}_{word} are defined as follows:

$$S^{train}_{word} = \{(I^{train}_i, L^{train}_i, W^{train}_i)\}^{N^{train}}_{i=1}$$

$$S^{test}_{word} = \{(I^{test}_j, W^{test}_j)\}^{N^{test}}_{j=1}$$

where N^{train} and N^{test} are the number of train and test samples, $I \in \mathbb{R}^{C \times H \times W}$ is an image when C, H, W are the image RGB channels, height and width, $L^s_i \in \{0, 1\}^{H \times W}$ is the label and W^s_i is the word vector of the class of the image I^s_i for $s \in \{train, test\}$.

3.2 ZSSS via Visual Class Embedding

In our proposed ZSSS via VCE setting, the training set S_{VCE}^{train} and the test set S_{VCE}^{test} exclude the word vector W as we generate the class embedding from the query image with ImageNet pretrained knowledge. The train set S_{VCE}^{train} and the test set S_{VCE}^{test} for our ZSSS via VCE setting are defined as follows:

$$S_{VCE}^{train} = \{(I_i^{train}, L_i^{train})\}_{i=1}^{N_{train}}$$

$$S_{VCE}^{test} = \{(I_j^{test})\}_{j=1}^{N_{test}}$$

Given such datasets, our task is to segment an image I into a mask $M \in \{0,1\}^{H \times W}$, meaning if a pixel belongs to the target class.

CHAPTER IV

Method

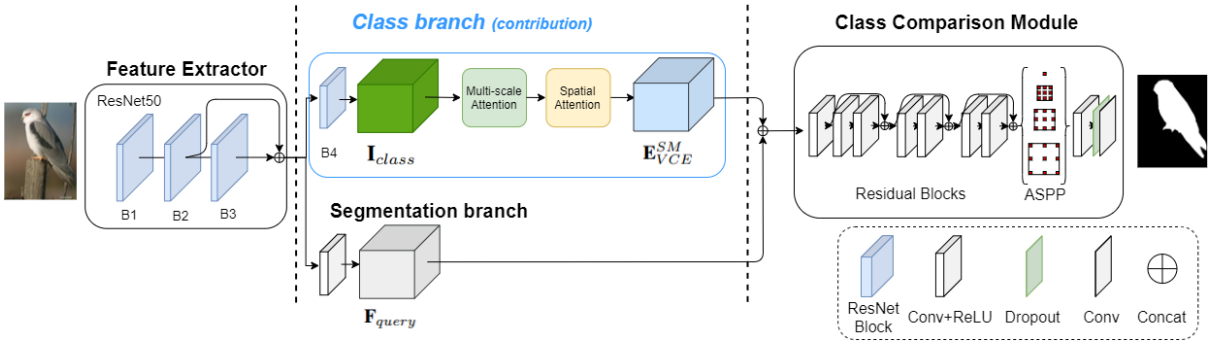

Figure 4.1: The overview of the proposed SM-VCENet, which consists of three parts: feature extractor, class and segmentation branches in parallel, and class comparison module. The feature extractor is shared for both the class branch and the segmentation branch. The details of the class branch are described in Figure 4.2.

In this section, we describe our motivation and the overall architecture of SM-VCENet.

4.1 Visual Class Embedding (VCE) Motivation

For predicting an unseen class, we need a class representative expression (class embedding) as we cannot have a classifier that solves the regression problem that if the query feature belongs to the unseen class. The class embedding can be in the form of constant vector, as a word is expressed in the form a

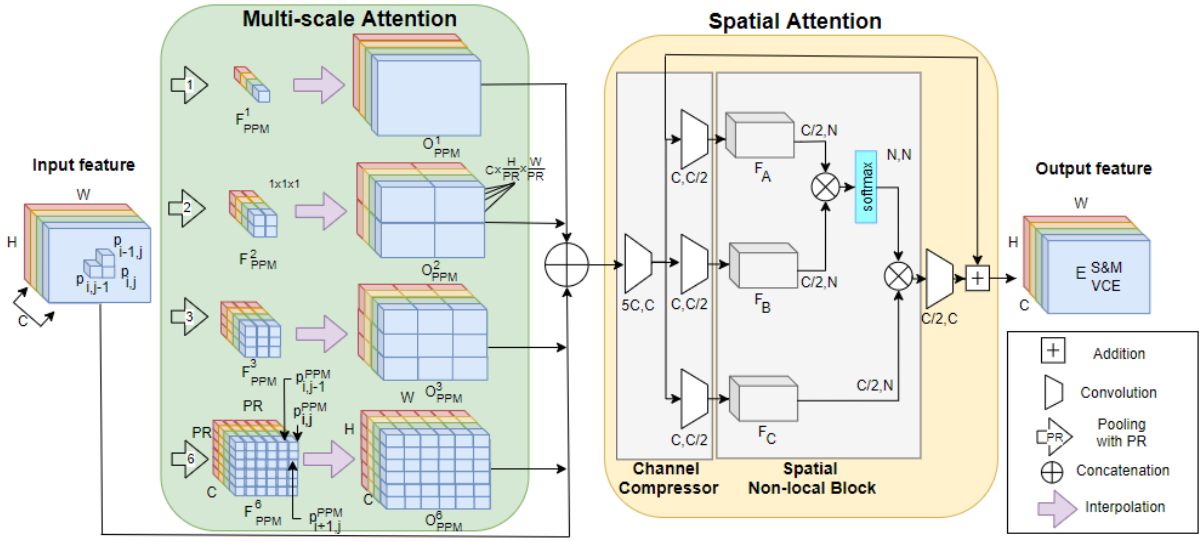

Figure 4.2: Multi-scale attention (MA) module and spatial attention (SA) module overview. The final output feature map from the modules has the same shape as the input feature. MA global average pools the input feature with various pooling rate and concatenates them. SA compresses the concatenated feature with convolution layers, then grasps spatial information by the non-local block.

vector from the word vector language model [16]. However the prior works [13–15] assumed the lingual knowledge can be directly used in the visual task, we think the two knowledge must form different distribution. Thus, instead, we propose to use ImageNet [23] pretrained knowledge for the class embedding. Shown in the Figure 1.1 (b), the well known image classification task is based on the class embedding generation by ResNet [37] and VGGNet [38]. Our proposed Visual Class Embedding (VCE) is from the query image with ImageNet pretrained networks. By freezing the pretrained backbone network, we preserve the visual domain knowledge.

While the class embedding from supporting images of FSSS are visual that includes visual knowledge (*e.g.*topological, spatial and scale information), the class embedding from the w-ZSSS approaches cannot contain such visual knowledge as a word vector cannot imply them. At the best of our knowledge, we are the first proposing visual knowledge oriented class embedding for zero-shot learning. Maximizing the advantages of visual knowledge, we propose multi-scale attention and spatial attention modules for visual class embedding.

4.2 Overall Architecture of SM-VCENet

SM-VCENet consists of three main parts: a feature extractor, class and segmentation branches, and class comparison module (CCM). Figure 4.1 illustrates the overall architecture of SM-VCENet. We use ImageNet-pretrained ResNet50 to extract the query image feature F_{query}. The class branch generates

visual class embedding(VCE). CCM conducts semantic segmentation by comparing F_{query} and VCE in pixel level with three residual blocks and ASPP [2].

We extract the query image feature by ImageNet pretrained ResNet50 [37] without updating the parameters. Given a set of ResNet blocks $\{\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$ and $\mathbf{B}_4\}$, the corresponding features $\mathbf{F}_1, \mathbf{F}_2$ and $\mathbf{F}_3$ are extracted from the RGB query image $\mathbf{X} \in \mathbf{R}^{3\times H_{input}\times W_{input}}$ as follows: $\mathbf{F}_i = \mathbf{B}_i(\mathbf{F}_{i-1})$ where $i \in 2, 3, 4$ and $\mathbf{F}_1 = B_1(X)$ where H_{input}, and W_{input} are the input image height, and width. Inspired by the previous study [20] of an output feature from ResNet [37] backbone network for FSSS, we compute the query features $\mathbf{F}_{query}$ for the segmentation branch by concatenating features $\mathbf{F}_2$ and $\mathbf{F}_3$, and the input of class branch from $\mathbf{B}_4$:

$$\mathbf{I}_{class} = \mathbf{B}_4(\mathbf{F}_2 \oplus \mathbf{F}_3) \tag{IV.1}$$

$$\mathbf{F}_{query} = f(\mathbf{F}_2 \oplus \mathbf{F}_3) \tag{IV.2}$$

where f is the 3×3 convolution operation.

4.2.1 Class branch

The class branch creates class embedding that contains both spatial information and multiple size of compressed features for predicting multi-scale objects throughout two modules: multi-scale attention (MA) module and spatial attention (SA) module. Figure 4.2 represents the overall operations for MA and SA.

Multi-scale attention module extracts implicit information of features of various sizes through pooling with multiple ratios. Inspired by PSPNet [1], we adopt the early stage of pyramid spatial pooling module to compact multi-scaled information. Given an input $\mathbf{I}_{class} \in \mathbb{R}^{C\times H\times W}$, let $p_{c,i,j} \in \mathbf{I}_{class}$ the each pixel value where the c, i, j represent the position of it in the three-dimensional input feature. MA compresses it by the pooling with pooling ratio of (1, 2, 3, 6). Each feature maps pooled with PR pooling ratio divide the $H \times W$ size feature map into the PR^2 number of regions and extract one representative values for each region. Each compressed value is calculated by averaging pixel values in the specific region:

$$p^{PPM}_{c,i,j} = \frac{\sum_{h=1}^{i+i*(H/NR-1)} \sum_{w=1}^{j+j*(W/NR-1)} p_{c,i,j}}{NR} \tag{IV.3}$$

where number of regions $NR = \frac{H\times W}{PR^2}$ and $0 < i, j < PR$. Then we have PR^2 sized of two-dimensional output for each pooling ratio. The output feature map with PR pooling ratio is calculated as follows:

$$\mathbf{F}^{PR}_{PPM} = \sum_{c=1}^{C} \sum_{w=1}^{PR} \sum_{h=1}^{PR} p_{c,h,w} \tag{IV.4}$$

We have four different size of outputs $\mathbf{F}^1_{PPM}, \mathbf{F}^2_{PPM}, \mathbf{F}^3_{PPM},$ and $\mathbf{F}^6_{PPM}$ with the pooling ratios in (1, 2, 3, 6). Note that each $p_{c,i,j}$ in $\mathbf{F}^{PR}_{PPM}$ represents the distinct compressed information of input feature map which helps the model to predict multi-scaled objects. Therefore, MA module considers totally $1 \times 1 + 2 \times 2 + 3 \times 3 + 6 \times 6 = 50$ number of compressed information from multi-scaled features. We expand all the outputs, $\mathbf{F}^i_{PPM} \in \mathbb{R}^{C \times PR \times PR}, i \in 1, 2, 3, 6$ to the $\mathbf{O}^i_{PPM} \in \mathbb{R}^{C \times H \times W}$ where $i \in 1, 2, 3, 6$. We concatenate them with the input feature $\mathbf{I}_{class}$. Therefore, we have the final output of multi-scale attention module as follows:

$$\mathbf{O}_M = \text{Cat}(\mathbf{O}^1_{PPM}, \mathbf{O}^2_{PPM}, \mathbf{O}^3_{PPM}, \mathbf{O}^6_{PPM}, \mathbf{I}_{class}) \tag{IV.5}$$

Spatial Attention (SA) module uses the MA output that combines feature information of different sizes to interlink dependencies of each multi-scaled information. Spatial attention module calculates related information of those having spatial information of different sizes through non-local [31] operations. Compressor in spatial attention module first compressed $\mathbf{O}_M \in \mathbb{R}^{5C \times H \times W}$ to the dimension of $C/2 \times H \times W$ for the effective non-local operations by two stages of convolution. For the first stage, we decrease the channel of feature from 5C to C. We separate the output of the first stage into three branches and compute convolutions for each features with distinct weight matrices. To do the matrix multiplication between the three-dimensional features in order to connect dense relationships, we first view the dimension of each input from $C \times H \times W$ into $C \times N$ where $N = H \times W$. We multiply the feature $\mathbf{F}_A \in \mathbb{R}^{C \times N}$ and transposed feature $\mathbf{F}_B \in \mathbb{R}^{N \times C}$ to output $\mathbf{F}_D$. The second multiplication output $\mathbf{F}_M$ is computed by the softmax output of $\mathbf{F}_D$ and $\mathbf{F}_C$. Note that all features in the non-local operation ($\mathbf{F}_A$, $\mathbf{F}_B$, $\mathbf{F}_C$, $\mathbf{F}_D$, and output $\mathbf{F}_M$) represent the condensed multi-scale information of feature. Therefore, two multiplications of each feature strengthen connections for global context between multi-scale information. We recover the reduced channel of $\mathbf{F}_M$ from $C/2$ to the C. The final output of Spatial non-local block is the concatenation of $\mathbf{O}_M$ and $\mathbf{F}_M$ as follows: $\mathbf{E}^{SM}_{VCE} = Concat(\mathbf{F}_M, f(\mathbf{O}_M))$

4.2.2 Class comparison module (CCM)

CCM performs semantic segmentation by comparing the class embedding E^{SM}_{VCE} and the query feature F_{query}, acting like the segmentation decoder in [2]. Given the two features, it concatenates all the vectors in pixel-wise, which preserves the mutual location information. For efficient implementation, we reduce the number of the channel of the concatenated feature from $256 + C$ to 256 with a 1×1 convolution layer. Next, CCM solves a regression problem that if a feature vector of a pixel of F_{query} belongs to the same class. With the following three sequential basic residual blocks [37] where a residual block consists of two 3×3 convolutional layers with skip connection, CCM compares the two features. Finally, after ASPP [2], CCM results a predicted binary mask $M \in \{0, 1\}^{H_{input} \times W_{input}}$.

Note that each convolutional layers in CCM are followed by ReLU activation function without batch normalization [39]. However, in zero-(few-)shot learning setting where each sample of a mini-batch may have a different target class, the layers from different batches may have different distribution expressing their own target class. Thus, normalizing a layer with respect to batch smooths the expressiveness of the layer about the target class.

CHAPTER V

Experiment

In this section, we demonstrate three zero-shot semantic segmentation experiments on the public benchmarks PASCAL VOC 2012 (PASCAL-5^i) [12] and MSCOCO 2017 (COCO-20^i) [11]. First, we trained models on the PASCAL training set and evaluate on the PASCAL test set to show recognition performances on the train domain. Second, we tested the PASCAL-trained models on the COCO test set to see generalization ability and spatial & multi-scale understanding of the models. Lastly, we conducted ablation studies about spatial attention and multi-scale attention.

5.1 Implementation Details

All experiments are conducted with PyTorch [40] framework, following the settings in [14, 20]. We employ the mean of cross entropy loss over all spatial locations in the output feature map. Models are optimized by Stochastic Gradient Descent (SGD) with mini-batches optimizer through 200 epochs on PASCAL-5^i train set. We set the initial learning rate to 0.0025, momentum as 0.9 and weight decay as 0.0005. All baselines and our SM-VCENet share ResNet50 [37] backbone.

5.1.1 Word vector

We used 300-dimensional word embedding vectors of GloVe [16] pretrained on Common Crawl with 840B tokens. Following ZSSS settings [14, 15], for the word embeddings of the classes expressed in multiple words in both PASCAL-5^i and COCO-20^i classes, we averaged the word vectors (word vectors

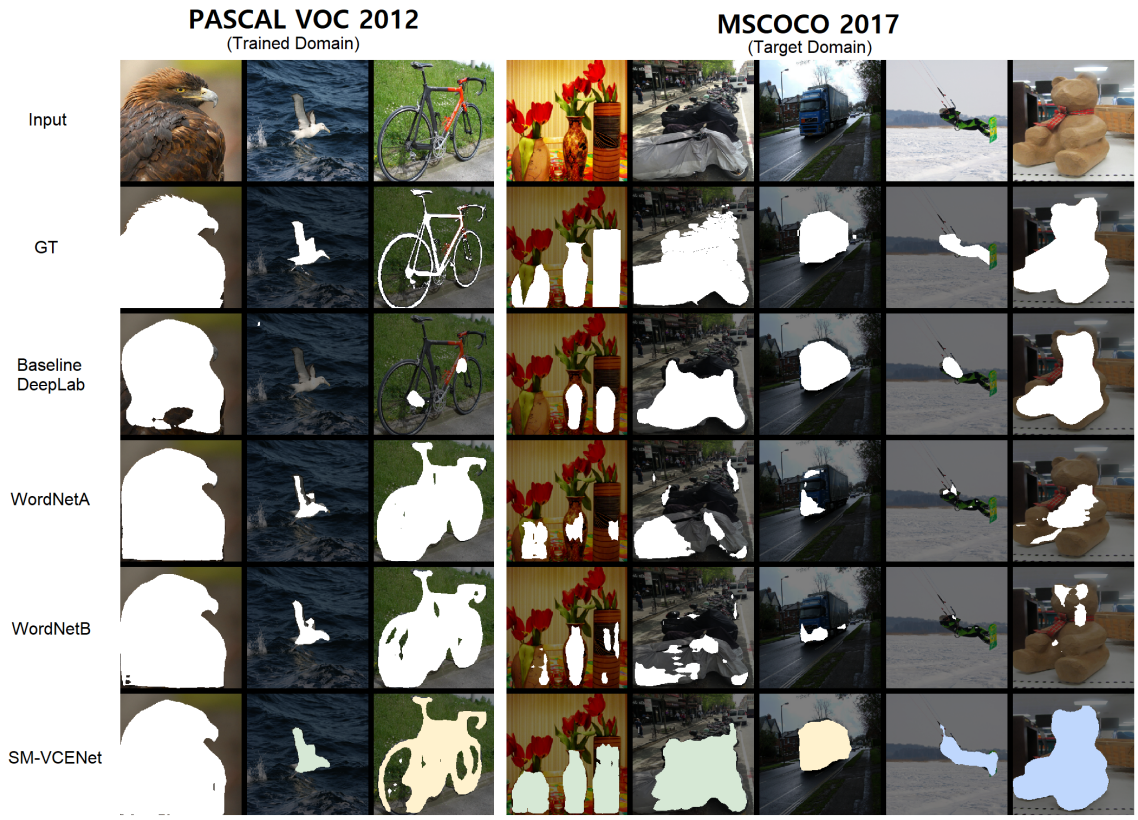

Figure 5.1: Qualitative comparisons of zero-shot semantic segmentation on PASCAL VOC 2012 and MS COCO 2017 test sets. GT denotes ground truth. We highlighted the strengths of our SM-VCENet: multi-scale understanding(light green), spatial understanding(yellow) and generalization(blue).

of "potted plant" and "tv/monitor" are the mean vector of each word in the label name) or simplified label name ("dinningtable" as "table").

5.1.2 Baselines

We set three baselines *Baseline-DeepLab*, *WordNet-A* and *WordNet-B*. Baseline-DeepLab is the same model as [2], which is designed for fully-supervised semantic segmentation. WordNet-A and WordNet-B are our re-implementation of w-ZSSS [14, 15], respectively, with Baseline-DeepLab backbone. Our reimplemented baselines are validated in Table 5.2, outperforming state-of-the-art approaches in PASCAL-5^i test set.

5.1.3 Evaluation metric

We adopted the standard evaluation metric, mean intersection-over-union (mIoU) as follows.

$$IoU = \frac{GT \cap Pred}{GT \cup Pred} = \frac{TP}{FN + TP + FP} \tag{V.1}$$

where *GT*, *Pred*, *TP*, *FP*, and *FN* are groud truth, prediction, true positive, false positive, and false negative, respectively. Note that all accuracy in our experiments are expressed in mIoU.

5.2 PASCAL-5^i (Trained domain)

The PASCAL-5^i [25] dataset is composed of images from PASCAL VOC 2012 [12] and additional annotations from SDS [41]. It has 41,040 training images and 4,000 test images. Shown in Table 5.1, It consists of four sub-datasets, dividing the 20 object classes in PASCAL. Each sub-dataset contains 15 training (seen) classes 5 test (unseen) classes. In order to directly compare ZSSS approaches to FSSS

Dataset	Unseen classes
PASCAL-5^0	aeroplane, bicycle, bird, boat, bottle
PASCAL-5^1	bus, car, cat, chair, cow
PASCAL-5^2	diningtable, dog, horse, motorbike, person
PASCAL-5^3	potted plant, sheep, sofa, train, tv/monitor

Table 5.1: Unseen classes for a four split cross-validation test on PASCAL-5^i dataset

approaches, we followed the same PASCAL-5^i FSSS setting [17, 18, 20, 21, 35] without using the paired supporting image.

5.2.1 Quantitative results

In Table 5.2, we provide fair comparisons between our SM-VCENet, the baselines, a ZSSS SOTA (ZSVM [14]) and one-shot semantic segmentation (OSSS) approaches [17, 18, 20, 21, 35]. Our implementations including SM-VCENet, WordNet-A and WordNet-B outperform the ZSSS SOTA (ZSVM) with about 10% relative margin in mIoU accuracy and achieve higher accuracy than an OSSS approach (Siam *et al.*, 2019 (AMP) [21]). Even further, WordNet-B achieves 51.0% mIoU which is the highest accuracy in ZSSS on PASCAL-5^i benchmark, reaching close to 56.0% mIoU score of OSSS SOTA (PGNet [18]).

Methods	Shot	Word	P-5^0	P-5^1	P-5^2	P-5^3	mIoU
Baseline-DeepLab	0		36.5	47.9	39.8	30.3	38.6
ZSVM [14]	0	✓	39.6	52.6	41.0	35.6	42.2
AMP-1 [21]	1		37.4	50.9	46.5	34.8	42.4
AMP-2 [21]	1		41.9	50.2	46.7	34.7	43.4
WordNet-A	0	✓	48.1	60.7	37.5	38.9	46.3
SM-VCENet	0		48.1	54.2	43.1	40.2	46.5
PANet [35]	1		42.3	58.0	51.1	41.2	48.1
WordNet-B	0	✓	49.6	66.3	46.3	41.7	51.0
PPNet [17]	1		52.7	62.8	**57.4**	47.7	55.2
CANet [20]	1		52.5	65.9	51.3	**51.9**	55.4
PGNet [18]	1		**56.0**	**66.9**	50.6	50.4	**56.0**

Table 5.2: Performance comparison of zero-shot and one-shot approaches on PASCAL-5^i test set. Shot denote the number of support image. The 'Word' column remarks if word-vector is supported. P-5^i denotes PASCAL-5^i test set.

5.2.2 Qualitative results

Figure 5.1 qualitatively compares the performances of SM-VCENet and baselines with PASCAL test set images. We prepared the three PASCAL-representative examples containing: *single and large* (the first column, big bird), *single and small* (the second column, small bird), and *single and delicate* (the third column, bicycle) object. While baseline-DeepLab works only on *single and large* case, both SM-VCENet and WordNet-A&B work well on the two bird cases, showing multi-scale awareness. However, for the bicycle sample, only our SM-VCENet recognize the fine wheel of the bike.

5.3 COCO-20^i (Target domain)

COCO-20^i consists of 12,468 test images from MS COCO 2017 [11] with 80 classes. Compared to the COCO-20^i setting in FSSS that requires pairing (supporting image-query image) and evaluates each test images more than once, our COCO-20^i setting for ZSSS evaluate every test images only once as the pairing is not necessary. We conducted four sub-datasets with 60 train classes and 20 test classes. We followed the same class splits of the sub-datasets from [22, 35]. (see Table 5.3).

5.3.1 Motivation

We present a unified ZSSS benchmark (PASCAL2COCO) with two advantages: it contains the *domain adaptation (generalization)* problem; it has challenging samples including *multi-scaled multiple* objects and *complex objects in noisy backgrounds*. First, existing ZSSS benchmarks [13, 15] cannot evaluate generalization ability as they train and test set on the same domain. However, in ZSSS setting where

COCO-20^0		COCO-20^1		COCO-20^2		COCO-20^3	
1	Person	2	Bicycle	3	Car	4	Motorcycle
5	Airplane	6	Bus	7	Train	8	Truck
9	Boat	10	T.light	11	Fire H.	12	Stop
13	Park meter	14	Bench	15	Bird	16	Cat
17	Dog	18	Horse	19	Sheep	20	Cow
21	Elephant	22	Bear	23	Zebra	24	Giraffe
25	Backpack	26	Umbrella	27	Handbag	28	Tie
29	Suitcase	30	Frisbee	31	Skis	32	Snowboard
33	Sports ball	34	Kite	35	B.bat	36	B. glove
37	Skateboard	38	Surfboard	39	T. racket	40	Bottle
41	W. glass	42	Cup	43	Fork	44	Knife
45	Spoon	46	Bowl	47	Banana	48	Apple
49	Sandwich	50	Orange	51	Broccoli	52	Carrot
53	Hot dog	54	Pizza	55	Donut	56	Cake
57	Chair	58	Couch	59	P. plant	60	Bed
61	D. table	62	Toilet	63	TV	64	Laptop
65	Mouse	66	Remote	67	Keyboard	68	Cellphone
69	Microwave	70	Oven	71	Toaster	72	Sink
73	Fridge	74	Book	75	Clock	76	Vase
77	Scissors	78	Teddy	79	Hairdrier	80	Toothbrush

Table 5.3: COCO-20^i category splits. For the i-th fold, the images of the 20 classes of the i-th split are used for evaluation, and the other images of the other splits are used for training.

no target domain information is given, measuring generalization ability is important. By conducting domain adaptation (from PASCAL to COCO), PASCAL2COCO benchmark can measure generalization ability. Second, tackling the difficulty of PASCAL-5^i dataset where almost images contain a single object occupying a large area, we propose to evaluate on the COCO-20^i dataset. COCO-20^i is more challenging dataset for ZSSS evaluation than PASCAL-5^i with the following two reasons: it contains 80 classes that is four times than which of PASCAL-5^i; it is a real-world scene containing different sized multiple objects (including small objects) in a single image or complex objects in noisy background.

5.3.2 Quantitative results

Table 5.4 shows performance comparisons between our SM-VCENet and the w-ZSSS approaches on PASCAL2COCO. We trained all models on the PASCAL-5^i splits $i \in \{0, 1, 2, 3\}$. We measured the accuracy of the models in the same domain of trained domain, PASCAL-5^i, and the different domain, COCO-20^i. WordNet-A&B and SM-VCENet achieve higher performance than Baseline-DeepLab in PASCAL-5^i test set. However, in the COCO-20^i test set, only SM-VCENet (41.34%) steadily outperform the baseline-DeepLab (30.53%). Although WordNet-A and WordNet-B achieve superior scores on

			PASCAL-5^i		COCO-20^i					
Models	Trained	Word	mIoU	Average	COCO-20^0	COCO-20^1	COCO-20^2	COCO-20^3	mIoU	Average
Baseline-DeepLab	PASCAL-5^0		36.49	38.59 (100%)	33.32	30.34	25.30	31.85	30.20	30.53 (100%)
	PASCAL-5^1		47.85		34.81	32.51	25.92	32.16	31.35	
	PASCAL-5^2		39.75		33.64	34.29	25.39	33.02	31.59	
	PASCAL-5^3		30.28		31.88	31.74	22.68	29.56	28.97	
WordNet-A	PASCAL-5^0	✓	48.07	46.32 (120.0%)	26.61	23.14	22.92	24.79	24.37	24.51 (80.3%)
	PASCAL-5^1	✓	60.73		30.13	29.55	23.45	22.23	26.34	
	PASCAL-5^2	✓	37.54		21.12	24.43	25.64	23.33	23.63	
	PASCAL-5^3	✓	38.92		26.57	22.07	20.99	25.23	23.72	
WordNet-B	PASCAL-5^0	✓	49.62	**50.95 (132.0%)**	40.94	36.08	30.44	32.09	34.89	31.65 (103.7%)
	PASCAL-5^1	✓	66.26		34.54	31.88	26.82	27.76	30.25	
	PASCAL-5^2	✓	46.28		37.00	37.90	33.07	30.64	34.65	
	PASCAL-5^3	✓	41.65		35.34	23.72	22.69	25.56	26.83	
SM-VCENet	PASCAL-5^0		48.09	46.51 (120.5%)	37.60	39.25	39.92	41.81	39.65	**41.34 (135.4%)**
	PASCAL-5^1		54.21		42.69	45.69	40.44	42.77	42.89	
	PASCAL-5^2		43.12		39.78	41.08	38.16	45.59	41.15	
	PASCAL-5^3		40.62		39.70	45.07	38.80	43.12	41.67	

Table 5.4: Zero-shot semantic segmentation performances including domain adaptation. Average column presents average mIoU score. % notation below the average mIoU score denotes the percentage score over the baseline-DeepLab. Models are trained on PASCAL-5^i and tested on both PASCAL-5^i and COCO-20^i test sets. Word column denotes if word vector is used for class embedding generation.

the PASCAL-5^i, their scores (24.51% and 31.65%) in COCO-20^i are below or slightly better than the baseline. In conclusion, (1) our SM-VCENet steadily outperforms the baseline in both the trained domain and the target domain, recognizing a novel class better; (2) w-ZSSS are overfit to the train domain because they failed to generalize and could not recognize an unseen class, performing similar to or less than the baseline.

5.3.3 Qualitative results

Figure 5.1 contains three challenging cases: one (the fourth and the fifth column), multiple objects in different scales; two (the sixth column), complex object in a noisy background; three (the seventh and the eighth column), unseen but simple to generalize objects.

Multi-scale Attention Shown in the fourth column and fifth column in Figure 5.1, only SM-VCENet accurately recognizes multi-scale multiple objects in a single image. Compared to the PASCAL-5^i dataset, COCO-20^i includes multiple objects appearing in various scale in an image. SM-VCENet recognizes the objects precisely through *multi-scale* attention module. Performing worse than the baseline-DeepLab, w-ZSSS approaches fail to classify multiple or multi-scale objects.

Spatial Attention The sixth column shows the example in which the appearance of the truck is very similar to the road and the trees nearby. Only SM-VCENet accurately predicts the segmentation labels for the truck in noisy background. WordNet series that lack spatial understanding fail in predicting the complex object with noisy background which requires understanding the global context.

Generalization The seventh and the eighth column in Figure 5.1 are about simple examples of 'person' and 'bear' categories. 'Person' category is included in the training classes, but the 'bear' class is not

included. SM-VCENet predicts the seen but different domain (background) and the unseen but simple cases clearly. However the both w-ZSSS approaches fail to recognize even the seen class *person* during training neither the unseen but simple *bear* case. either trained class or not, meaning w-ZSSS approaches cause to overfit to the trained domain.

5.4 Ablation Study for Multi-Scale and Spatial Attention

Table 5.5 represents the efficiency of multi-scale attention and spatial attention. VCENet refers to a model with the visual class branch using global average pooling instead of multi-scale attention. The M and S columns of the table represent the presence of multi-scale attention and spatial attention modules, respectively. On the PASCAL-5^i test set (trained domain), MA and SA improve the accuracy from 45.01% to 45.77% and 46.51%, each time when MA and SA are added to the VCENet. On the COCO-20^i test set (target domain), the accuracy of M-VCENet and SM-VCENet is increased from 34.43% to 40.81% and 41.34%, showing equally high accuracy in the unseen domain, COCO-20^i. Figure 5.2 shows the qualitative results of the ablation study from COCO-20^i test set image. Compared to VCENet, multi-scale attention module (M-VCENet) recognizes the target class in various sizes better. Furthermore, SM-VCENet recognizes, which includes spatial attention module, achieves higher accuracy and clear prediction.

	M	S	P-mIOU	C-mIOU
Baseline-Deeplab			38.59	30.53
VCENet			45.01	34.43
M-VCENet	✓		45.77	40.81
SM-VCENet	✓	✓	46.51	41.34

Table 5.5: Ablation study for the multi-scale and spatial attention module.

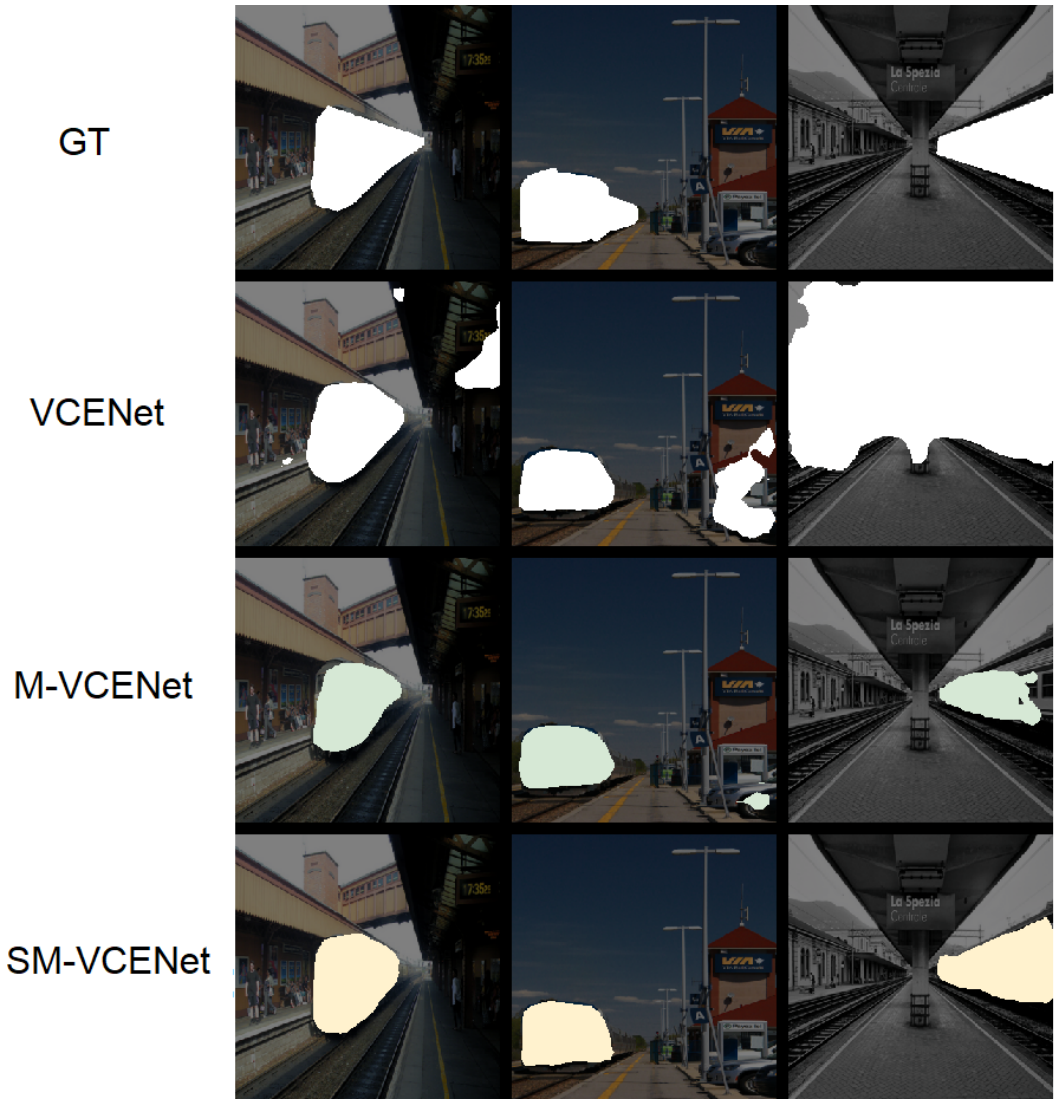

Figure 5.2: Ablation study about multi-scale attention module and spatial attention module. GT denotes the ground truth. VCENet is the visual class branch included DeepLab. M-VCENet is multi-scale attention module added VCENet. SM-VCENet is VCENet including spatial attention and multi-scale attention modules.

CHAPTER VI

Conclusion

Fully supervised semantic segmentation methodologies using large-scale dataset have made rapid progress. However, such methods burden expensive labeling cost of the large data. To solve it, recent zero-shot semantic segmentation approaches are proposed, which is to recognize any unseen class. In this paper, we propose SM-VCENet for zero-shot semantic segmentation that achieves domain-aware visual class embedding from the ImageNet pretrained knowledge and the query image, tackling the domain-agnostic class embedding of w-ZSSS that is constrained to the language-based word vector. Our SM-VCENet achieves significant robustness in generalization and performance on challenging visual scene understanding while w-ZSSS fails. Moverover, multi-scale and spatial attention modules in our model enable to predict multi-scaled multiple objects in the scenes and complex objects in a noisy background. Lastly, to evaluate the generalization ability and performance on the more challenging scens in the real-world, we proposed the novel benchmark (PASCAL2COCO).

In the future, the knowledge distribution gap between the vision field and the language field should be researched if one wants to utilize the lingual knowledge for a visual task in practice.

References

[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890. 11, 14, 20

[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. 11, 20, 21, 24

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818. 11

[4] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen, "Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12475–12485. 11

[5] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al., "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020. 11

[6] Ping Chao, Chao-Yang Kao, Yu-Shan Ruan, Chien-Hsiang Huang, and Youn-Long Lin, "Hardnet: A low memory traffic network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3552–3561. 11

[7] Marin Orsic, Ivan Kreso, Petra Bevandic, and Sinisa Segvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 12607–12616. 11

[8] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *arXiv preprint arXiv:2004.02147*, 2020. 11

[9] Juntang Zhuang, Junlin Yang, Lin Gu, and Nicha Dvornek, "Shelfnet for fast semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0. 11

[10] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4151–4160. 11

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755. 11, 23, 26

[12] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015. 11, 23, 25

[13] Maxime Bucher, VU Tuan-Hung, Matthieu Cord, and Patrick Pérez, "Zero-shot semantic segmentation," in *Advances in Neural Information Processing Systems*, 2019, pp. 468–479. 11, 15, 19, 26

[14] Naoki Kato, Toshihiko Yamasaki, and Kiyoharu Aizawa, "Zero-shot semantic segmentation via variational mapping," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0. 11, 13, 15, 16, 19, 23, 24, 25, 26

[15] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8256–8265. 11, 15, 16, 19, 23, 24, 26

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543. 11, 19, 23

[17] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He, "Part-aware prototype network for few-shot semantic segmentation," *ECCV*, 2020. 11, 15, 25, 26

[18] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao, "Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9587–9595. 11, 15, 25, 26

[19] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu, "Crnet: Cross-reference networks for few-shot segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4165–4173. 11, 15

[20] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen, "Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5217–5226. 11, 15, 20, 23, 25, 26

[21] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand, "Amp: Adaptive masked proxies for few-shot segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5249–5258. 11, 15, 25, 26

[22] Khoi Nguyen and Sinisa Todorovic, "Feature weighting and boosting for few-shot segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 11, 15, 26

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015. 12, 19

[24] Antonio Torralba and Alexei A Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528. 12

[25] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots, "One-shot learning for semantic segmentation," *BMVC*, 2017. 13, 25

[26] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. 14

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015. 14

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440. 14

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao, "Deep high-resolution representation learning for visual recognition," *TPAMI*, 2019. 14

[30] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia, "PSANet: Point-wise spatial attention network for scene parsing," in *ECCV*, 2018. 14

[31] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803. 14, 15, 21

[32] Kaiyu Yue, Ming Sun, Yuchen Yuan, Feng Zhou, Errui Ding, and Fuxin Xu, "Compact generalized non-local network," in *Advances in Neural Information Processing Systems*, 2018, pp. 6510–6519. 14

[33] Zhen Zhu, Mengde Xu, Song Bai, Tengteng Huang, and Xiang Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 593–602. 14

[34] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012. 15

[35] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng, "Panet: Few-shot image semantic segmentation with prototype alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 15, 25, 26

[36] Jinghua Wang and Jianmin Jiang, "Adversarial learning for zero-shot domain adaptation," *ECCV*, 2020. 15

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 19, 20, 21, 23

[38] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 19

[39] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015. 22

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019. 23

[41] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," in *European Conference on Computer Vision*. Springer, 2014, pp. 297–312. 25

CHAPTER VII

Acknowledgement

Special thanks to Yooseung Wang from Agency for Defense Development, Daejeon, Korea for providing inspirations and experimental results about Spatial attention module and Multi-scale attention module.